Transforming Query and Data Retrieval Systems With the Advanced Power of GPT-4o: Generative AI at the Forefront of Extracting Data

MSR167

Rajdeep Kaur[1], Vedant Soni[1], Nicola Waddell[2], Shubhram Pandey[1], Gagandeep Kaur[1], Barinder Singh[2]
[1]Pharmacoevidence, Mohali, India, [2]Pharmacoevidence, London, UK

CONCLUSION

The AI-driven query and data retrieval interface has the ability to revolutionize the way researchers extract and analyze data from PDFs and other document formats. This innovative solution significantly streamlines information retrieval processes in healthcare economics research, allowing efficient access to crucial insights while optimizing resource utilization. The system's validated performance shows its potential for accelerating research workflows and supporting evidence-based decision-making.

INTRODUCTION

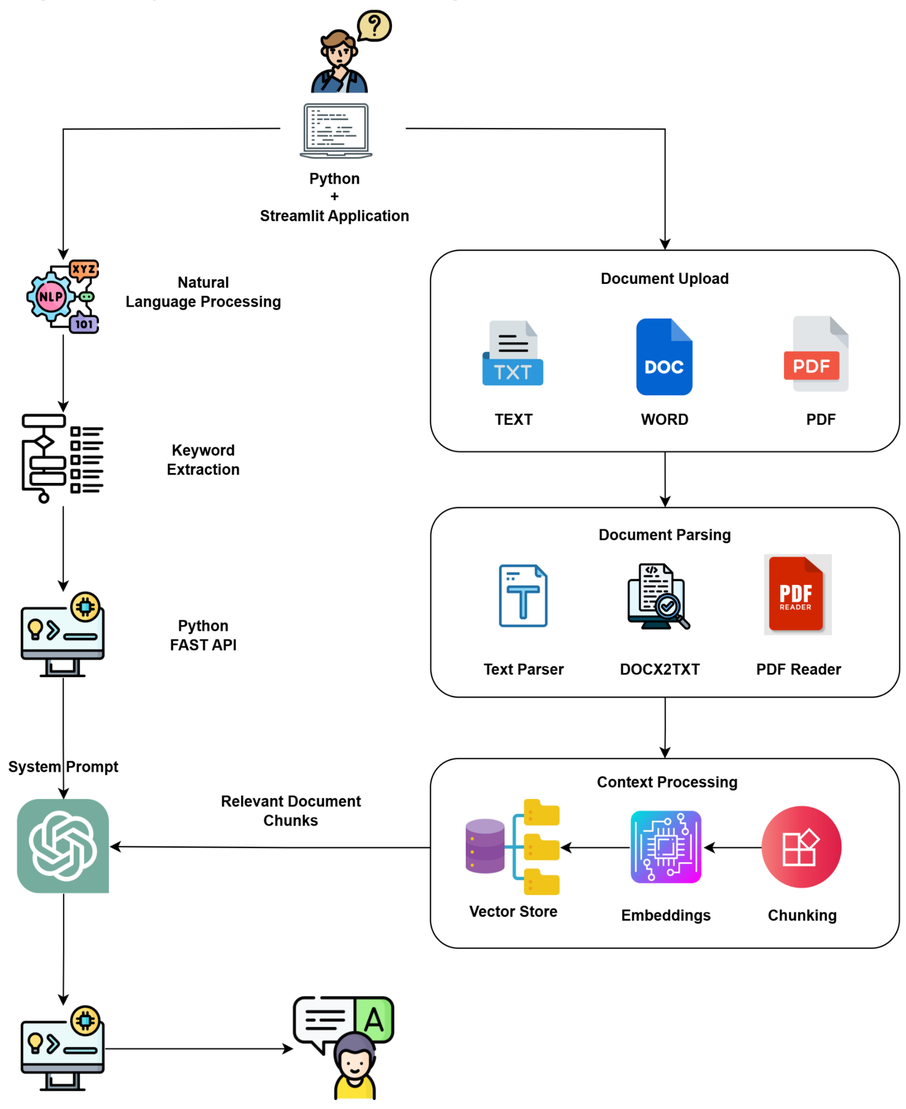
- The emergence of large language models (LLMs) has resulted in unprecedented opportunities to revolutionize query and data retrieval systems due to the advancements in generative artificial intelligence
- This research demonstrates an innovative automated data extraction interface leveraging GPT-4o and OpenAI Embeddings, designed specifically for precise information retrieval from research articles
- The system deploys advanced chunking mechanisms and FAISS indexing for enhanced similarity-based searches, facilitating contextually relevant responses
- The section-focused methodology and context-aware processing significantly enhance the quality of responses while simultaneously reducing hallucinations

OBJECTIVE

To automate the data extraction and information retrieval from research articles by developing a query interface. The interface integrated the GPT-4o and OpenAI embeddings to enable intelligent data extraction. This aimed to :

1. Develop an efficient document processing pipeline starting with an effective chunking algorithm (Divide documents into context-aware pieces), followed by FAISS indexing for fast information retrieval as shown in Figure 1
2. To respond to user enquiries by retrieving relevant information from the document content

Figure 1: Systematic workflow diagram for PDF extraction



METHODS

Interface

- The development of the data extraction interface was based on state-of-the-art language models that are extended to sufficiently index, as shown in Figure 2
- The architecture builds on GPT-4o, integrated with the Python programming language, and deployed using Streamlit

Preprocessing pipeline

- The document processing pipeline used a sophisticated chunking algorithm for systematically dividing uploaded documents into optimally sized chunks that can be processed efficiently while maintaining awareness of the context

METHODS (contd.)

- These chunks were vectorized using OpenAI's embedding technology to convert them into high-dimensional representations that capture semantic relationships. FAISS (Facebook AI Similarity Search) gave a solution that is scalable for similarity-based search, yielding quick content retrieval

Prompt Development

- The Prompt was developed using the section focus technique, where the model knows from which section of the document the question has been asked and only focuses on those sections of the documents
  Sample prompts are:
  - "What are the key findings of the study?“
  - "What was the proportion of males and females in this study?“
  - "Give me results of the main outcomes being assessed in this trial?“

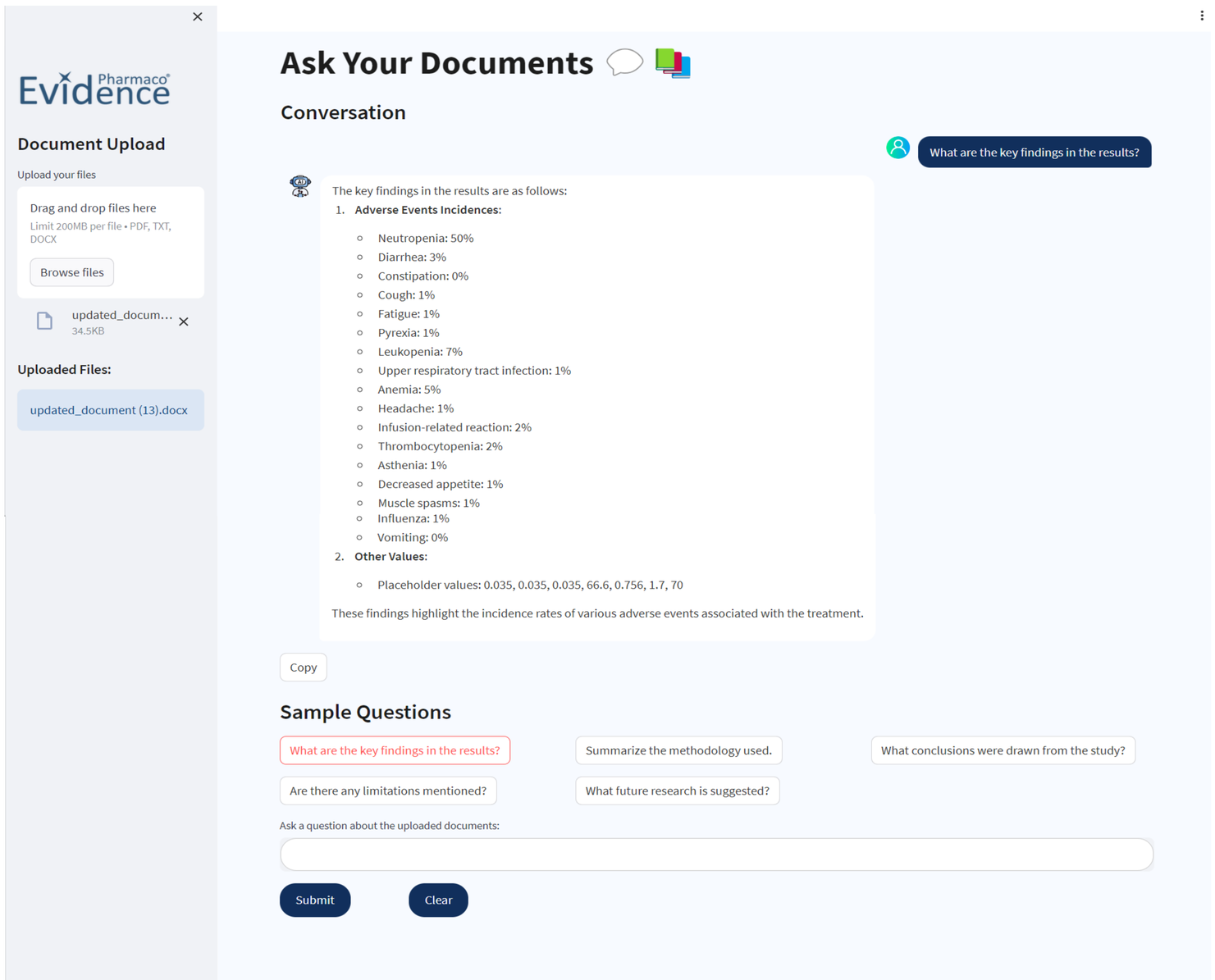
Relevant Information Retrieval

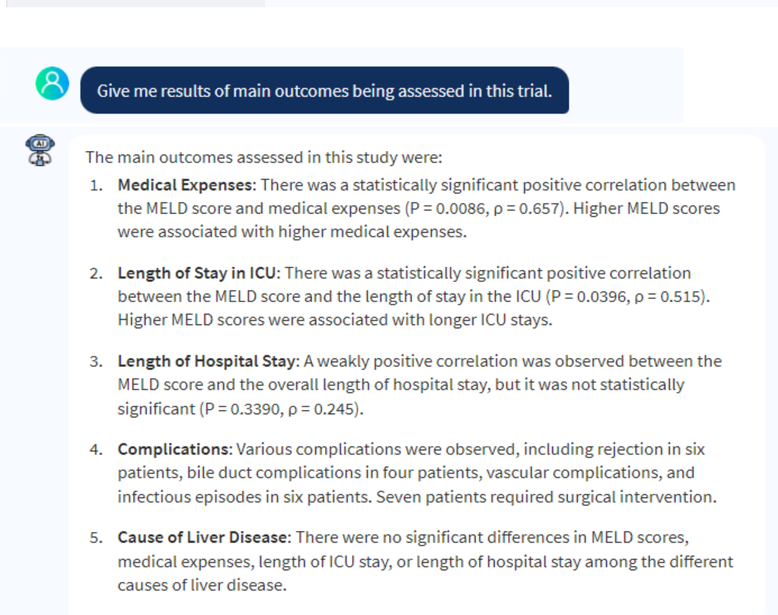
- The context from each section was combined to generate a unified answer. This unified answer was pushed to the frontend and shown to the user

RESULTS

- Using a mixed test set with 56 prompts on four different research publications, the system performed competently to retrieve information accurately
- Domain experts extensively tested the system's responses, finding satisfactory performance with correct data retrieval for 50 of 56 prompts
- Though inconsistencies in 6 prompts indicated a need for further optimization for complex queries

Figure 2: Web interface for PDF Extraction





- The qualitative feedback from experts suggested the system performed well in demonstrating context coherence, producing well-structured responses
- A predefined question has been asked to the model, and based on the section focus as described in the prompt, it picks up the section and only focuses on those specific sections of the documents
- The model only selected those sections of the document for the information retrieval that match the user query.
- For faster response and less hallucination in the model, section focus is necessary, and it yields a higher-quality response

References

1. Bolanos et al. *arXiv preprint*. 2024
2. de la Torre-López et al. *Computing*. 2023;105(10):2171-2194
3. Jonnalagadda et al. *Syst Rev*. 2015;4:1-16

Disclosures

RK, BS, VS, NW, SP, and GK authors, declare that they have no conflict of interest

Need a poster copy?

Scan here

Hands-on experience with AI tools?

Contact: rshiny@pharmacoevidence.com